# Chapter 8

# Transcriptional Profiling of *Dictyostelium* with RNA Sequencing

**Edward Roshan Miranda*, Gregor Rot*, Marko Toplak, Balaji Santhanam, Tomaz Curk, Gad Shaulsky, and Blaz Zupan**

## Abstract

Transcriptional profiling methods have been utilized in the analysis of various biological processes in *Dictyostelium*. Recent advances in high-throughput sequencing have increased the resolution and the dynamic range of transcriptional profiling. Here we describe the utility of RNA sequencing with the Illumina technology for production of transcriptional profiles. We also describe methods for data mapping and storage as well as common and specialized tools for data analysis, both online and offline.

**Key words** *Dictyostelium*, RNA sequencing, Multiplexing, Web-based applications, Visual programming, Data mining, Differential expression, Orange, dictyExpress, PIPA

## 1 Introduction

In the past decade, a significant understanding of dictyostelid transcriptomes has been achieved, thanks to techniques such as rapid amplification of cDNA ends (RACE), Sanger sequencing of cDNAs, and microarrays (1–4). The recent development of RNA sequencing (RNAseq) has lead to further appreciation of the complexity of dictyostelid transcriptomes and to vast improvements in transcriptome quantification (5). RNAseq is a high-throughput method that employs massive parallel sequencing of cDNA fragments generated from RNA (6). The method generates millions of short sequencing reads that represent fragments of the transcriptome. These fragments are then mapped to the genome of interest or assembled de novo. The number of fragments that map to a specific gene is directly proportional to the abundance of the respective RNA in the sample. The large number of sequencing reads enables the landscaping of transcriptomes at unprecedented depth and resolution.

*Edward Roshan Miranda and Gregor Rot have contributed equally to this work.

Ludwig Eichinger and Francisco Rivero (eds.), *Dictyostelium discoideum Protocols*, Methods in Molecular Biology 983,
DOI 10.1007/978-1-62703-302-2_8, 

RNAseq has been used to improve existing gene models, including predicting exon–intron boundaries and untranslated regions, to identify alternative splicing of transcripts, and to discover new genes (7, 8). Determination of quantitative and qualitative changes in RNA is possible at a wide dynamic range. RNAseq has supplanted microarrays as the technique of choice for understanding genome wide expression patterns. It yields a digital output of RNA quantity, as opposed to the analog output of microarrays, and it is free of some microarray limitations, including variable hybridization kinetics and cross hybridization among different hybridization targets. Due to the high reproducibility of RNAseq, technical replications are no longer needed—only biological replications are required.

Next generation sequencing technologies have improved appreciably since their introduction, yielding improved read quality and quantity. Currently, each sequencing run yields more reads than needed for most applications, so multiplexing is employed as a means of cost reduction (9). In this chapter we describe the techniques of RNAseq, with and without multiplexing, using the Illumina platform.

mRNA accounts for about 2% of the total RNA in *Dictyostelium* cells so it must be enriched before the analysis. Here we describe a method that begins with the isolation of polyA$^+$ mRNA by hybridization to oligo dT beads. We describe the preparation of cDNA from the enriched mRNA and the preparation of either single-sample libraries or pools of samples with multiplexing.

Analysis of RNAseq data consists of deconvolution in the case of multiplexed data, mapping the reads to the genome, and processing the data into values that represent transcript abundance. We describe the process of data analysis and storage as well as several examples of downstream data analysis, such as differential gene expression.

## 2 Materials

### 2.1 Reagents

#### 2.1.1 RNA Purification and cDNA Synthesis

The reagents must be RNAse free. Use disposable sterile plasticware and clean the work areas and the pipettors with RNAseZap (Ambion) before each procedure. Always wear gloves, mask, and lab coat when handling RNA (see Note 1).

Water and aqueous solutions used for RNA work should be treated with diethylpyrocarbonate (DEPC) to inactivate RNAse. Add 0.1% DEPC to the solution, incubate overnight at room temperature, and autoclave (15–25 min, liquid cycle). Do not DEPC-treat solutions that contain Tris.

1. TRIzol® (Life Technologies).

2. 10× MOPS buffer: 0.1 M MOPS, 5 mM EDTA, 25 mM sodium acetate; adjust to pH 7.0 with acetic acid and treat with DEPC.
3. Dynabeads mRNA Purification Kit (Life Technologies) supplied with oligo(dT) beads, binding buffer, washing buffer, and 10 mM Tris–HCl.
4. 10× Fragmentation buffer (Ambion).
5. Stop buffer (Ambion).
6. Glycogen (Ambion): 5 μg/μL.
7. 3 M sodium acetate, pH 5.2, DEPC treated.
8. 100% and 70% ethanol.
9. Random hexamer primers (Invitrogen): 3 μg/μL.
10. 100 mM dNTP set (Life Technologies).
11. 10 mM dNTP mix. Mix 10 μL of each dNTP from the 100 mM dNTP set and 60 μL of water.
12. RNaseOUT (Invitrogen): 40 U/μL.
13. SuperScript II (Invitrogen): 200 U/μL, supplied with 5× first-strand buffer and 100 mM DTT.
14. 10× second-strand buffer: 500 mM Tris–HCl, pH 7.8, 50 mM $MgCl_2$, 10 mM DTT.
15. RNaseH (Invitrogen): 2 U/μL.
16. *E. coli* DNA polymerase I (Invitrogen): 10 U/μL.
17. Microcentrifuge test tubes (1.5-mL, 0.5-mL, 2-mL) and sterile aerosol-resistant pipette tips (10-μL, 200-μL, 1-mL) (see Note 2).

#### 2.1.2 Single-Sample Library Preparation

1. Genomic DNA Sample Prep Kit (Illumina). Components of this kit can be replenished using the reagents mentioned below. Adapter oligonucleotides and PCR primers can also be ordered separately. Their sequences are available from the manufacturer.
2. 100 mM ATP (Sigma Aldrich): in water.
3. 10 mM dNTP mix. See Subheading 2.1.1, item 11.
4. T4 DNA polymerase (Invitrogen): 5 U/μL.
5. Klenow DNA polymerase (Invitrogen): 5 U/μL, supplied with 10× Klenow buffer.
6. T4 polynucleotide kinase (Invitrogen): 10 U/μL.
7. 1 mM dATP: dilute from the 100 mM dNTP set (see Subheading 2.1.1, item 10) in water.
8. DNA ligase (Invitrogen): 5 U/μL, supplied with 5× DNA ligase buffer.

9. 25 mM dNTPs: mix equal volumes of all four dNTPs from the 100 mM dNTP set (see Subheading 2.1.1, item 10).
10. Phusion DNA polymerase (New England BioLabs): 2 U/μL, supplied with 5× Phusion HF buffer.
11. QIAquick PCR spin kit (Qiagen) supplied with EB solution.
12. QIAquick MinElute kit (Qiagen) supplied with EB solution.
13. QIAquick gel extraction kit (Qiagen) supplied with EB solution.
14. 100 bp DNA ladder (Life Technologies).
15. Agarose (Calbiochem).
16. 50× TAE buffer: 242 g of Tris base, 57.1 mL of glacial acetic acid, 100 mL of 0.5 M EDTA, pH 8.0 in 1 L of water.
17. Ethidium bromide (Sigma), 10 mg/mL stock solution.
18. Bioanalyzer DNA 1000 chip (Agilent).

*2.1.3 Multiplexed Library Preparation*

1. Agencourt AMPure XP 60 mL Kit (Beckman Coulter). This kit includes carboxyl-coated magnetic beads.
2. 100% and 70% Ethanol.
3. Tween 20 (Fisher Scientific).
4. 10× Buffer Tango (Thermo Scientific).
5. 25 mM dNTPs (mix equal volumes of all four dNTPs from the 100 mM dNTP set; see Subheading 2.1.1, item 10).
6. 100 mM ATP (Sigma Aldrich) in water.
7. T4 DNA ligase (Fermentas): 5 U/μL, supplied with 10× T4 DNA ligase buffer and 50% PEG-4000 solution.
8. T4 DNA polymerase (Fermentas): 5 U/μL.
9. T4 polynucleotide kinase (Fermentas): 10 U/μL.
10. *Bst* DNA polymerase, large fragment (New England BioLabs) supplied with 10× ThermoPol reaction buffer.
11. Agarose (Calbiochem).
12. 50× TAE buffer: See Subheading 2.1.2, item 16.
13. Ethidium bromide: See Subheading 2.1.2, item 17.
14. Quantitative PCR kit with SYBRE green such as SYBR® Green PCR Master Mix (Life Technologies).
15. Phusion Hot Start High-Fidelity DNA Polymerase (New England BioLabs) supplied with 5× Phusion HF buffer.
16. 25 bp DNA Ladder (Life Technologies).
17. EB buffer, supplied with QIAquick PCR spin kit (Qiagen) or QIAquick gel extraction kit (Qiagen). This buffer can be prepared as 10 mM Tris–HCl, pH 8.5.

18. EBT: EB with 0.05% (v/v) Tween 20.
19. Oligonucleotide hybridization buffer: 500 mM NaCl, 10 mM Tris–HCl, pH 8.0, 1 mM EDTA in water.
20. Wide orifice pipette tips (VWR).
21. Hard-shell thin-walled 96-well skirted PCR plates for Quantitative-PCR (Bio-Rad).
22. Microseal "B" film PCR sealers (Bio-Rad).
23. Kit and reagents for DNA sequencing (Illumina).
24. Cluster generation kit (Illumina).
25. Multiplexing sequencing primer kit (Illumina). Alternatively, the following primers may be used for sequencing:
    (a) Read 1 sequencing primer: 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′
    (b) Index read sequencing primer: 5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′
    (c) Read 2 sequencing primer: 5′-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3′
26. Oligonucleotides for library preparation:

    Adapter_A1: A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T

    Adapter_A2: G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T

    Adapter_A3: A*G*A*T*CGGAA*G*A*G*C

    Primer_P1: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT

The sequences correspond to the order from 5′ to 3′ from left to right; * indicates a phosphothioate bond. All of the oligonucleotides should be ordered as HPLC purified and dissolved in water. Ask the supplier to synthesize and purify each primer in a separate batch to avoid cross contamination. Adapters A1, A2, and A3 are dissolved at 500 μM, and Primer_P1 at 10 μM. Order the primers in a 96-well plate to facilitate multichannel pipetting. The sequences of the primers, the criteria used to design them, and additional information are available in ref. (10).

### 2.2 Equipment

1. Two water incubators, one at 65°C and one at 80°C.
2. Heating blocks at different temperatures.
3. Agencourt SPRIPlate Super Magnet Plate (Beckman Coulter) for 96-well plates or DynaMag™-2 magnet (Life Technologies) for individual microcentrifuge tubes.

4. Rotisserie-style shaker/rotator with clamps for microcentrifuge test tubes (e.g., DiaMag Rotator, Diagenode).
5. Nanodrop spectrophotometer (Thermo scientific).
6. Thermal cycler such as PTC 100 (MJ Research) capable of holding 0.2-mL PCR test tubes or 0.5-mL test tubes.
7. Pipettors capable of dispensing 0.2 μL, 20 μL, 200 μL, and 1 mL (see Note 1).
8. Agarose gel electrophoresis equipment (e.g., Bio-Rad).
9. UV transilluminator (e.g., Kodak).
10. 96-well plate centrifuge (e.g., Eppendorf 5810R).
11. Microcentrifuge (e.g., Eppendorf 5415D).
12. Illumina Cluster Station.
13. Agilent 2100 Bioanalyzer (Agilent Technologies).
14. Real-time PCR machine (e.g., DNA engine Opticon 2, MJ Research).

### 2.3 Analysis Software

1. PIPA (http://pipa.biolab.si), a web-based tool for sequencing data management and bioinformatics analysis.
2. dictyExpress (http://dictyexpress.biolab.si), a web-based interactive gene expression analysis program.
3. Orange (http://pipa.biolab.si), a general purpose interactive data analysis environment.

## 3 Methods

### 3.1 RNA Purification and cDNA Synthesis

#### 3.1.1 Preparation of Total RNA

1. *Dictyostelium* cells are grown and developed under standard conditions (11) or as required by the desired experimental design.
2. Prepare total RNA using the TRIzol® reagent according to the manufacturer's recommendations (see Note 3).
3. Store the cell lysates in the TRIzol reagent at −80°C until all the samples are ready for the next step (see Note 4).
4. Dissolve the total RNA in 1× MOPS buffer.
5. Measure the RNA concentration using a spectrophotometer ($1AU_{260} = 40$ μg/μL).
6. Adjust the concentration to 1 μg/μL.
7. Store the total RNA samples in aliquots at −80°C. Do not thaw and refreeze the samples more than three times.

#### 3.1.2 mRNA Purification

mRNA isolation is performed using the Dynabeads mRNA Purification Kit from Life Technologies. Perform two rounds of

mRNA purification to ensure that more than 90% of the sequencing reads are from mRNA. Use the same aliquot of beads twice with an intermediate cleaning step to eliminate traces of sample from the first round. We recommend using 5–50 μg of total RNA as the starting material (see Note 5).

1. Put 10 μg of total RNA in a 1.5-mL RNAse-free microcentrifuge tube. Adjust the volume to 25 μL with DEPC-treated water.
2. Incubate the sample at 65°C for 5 min to disrupt secondary structures. Place the test tube on ice.
3. Aliquot 50 μL of Dynal oligo(dT) beads into a fresh 1.5-mL RNAse-free microcentrifuge tube.
4. Wash the beads twice with 50 μL of binding buffer. Place the microcentrifuge tube on the Dynal magnet and allow the beads to settle for 30 s. Once the supernatant is clear, remove it by pipetting with a plastic tip.
5. Resuspend the beads in 25 μL of binding buffer and add the 25 μL of total RNA from step 2. Rotate the tube at room temperature for 5 min, remove and discard the supernatant as described in step 4.
6. Wash the beads twice with 50 μL of washing buffer B as described in step 4.
7. Prepare for second round of purification by aliquoting 25 μL of binding buffer to a fresh 1.5-mL RNAse-free microcentrifuge tube.
8. Remove as much of the supernatant as possible from the beads of step 6. It is very important not to leave any supernatant in the test tube.
9. Add 25 μL of 10 mM Tris–HCl and incubate the samples at 80°C for 2 min to elute the mRNA. Immediately place the test tube in the Dynal magnet stand and transfer the supernatant (mRNA) to the test tube from step 7. Add 50 μL of washing buffer B to the remaining beads.
10. Incubate the mRNA sample from step 9 at 65°C for 5 min and place the test tube on ice.
11. Resuspend the beads from step 9 by finger flicking the test tube. Place the test tube on the Dynal magnet and remove the supernatant. Wash the beads once with 50 μL of binding buffer as in step 4 and remove the supernatant. Resuspend the beads in 25 μL of binding buffer.
12. Add 25 μL of the RNA sample from step 10 back into the tube from step 11. Rotate the test tube at room temperature for 5 min and discard the supernatant.
13. Wash the beads once with 50 μL of washing buffer B as in step 4 and remove the supernatant as in step 8.
14. Add 12 μL of 10 mM Tris–HCl and incubate the test tube at 80°C for 2 min to elute the mRNA. Immediately place the test

tube in the magnet stand and transfer the supernatant (mRNA) to a fresh microcentrifuge test tube.

15. Quantify the mRNA with a Nanodrop spectrophotometer (see Note 6).

Typically, 10 μg of total *Dictyostelium* RNA yield 100–200 ng of mRNA. Alternatively, one can start with 100 ng of mRNA if any other method of mRNA purification is used. Lower amounts of mRNA are also compatible with the next steps (see Note 7).

### 3.1.3 mRNA Fragmentation

mRNA fragmentation relies on metal ion-based catalysis and high temperature. Other protocols use heat alone, but we have observed that *Dictyostelium* mRNA is surprisingly stable at high temperatures, so we optimized the combination of chemical catalysis and high temperature to produce the desired fragment size of approximately 200 bases (see Note 8). We process 8 samples at one time for fragmentation and deal with any higher number in batches.

1. Start with 100 ng of purified mRNA (Subheading 3.1.2). Adjust the volume to 9 μL with water.
2. Add 1 μL of 10× fragmentation buffer and incubate at 70°C for 5 min.
3. Add 1 μL of stop buffer, mix by repeated pipetting, and place the test tube on ice.

### 3.1.4 Precipitation of Fragmented mRNA

1. Transfer 11 μL of the fragmented mRNA solution from Subheading 3.1.3 into an ice-cold 1.5-mL microcentrifuge test tube.
2. Add 1 μL of 3 M sodium acetate pH 5.2, 2 μL of glycogen (5 μg/μL) and 30 μL of 100% ethanol.
3. Mix by repeated pipetting and incubate at −80°C for 30 min.
4. Centrifuge at 18,000 × *g* in an Eppendorf centrifuge for 25 min at 4°C. A pellet should be visible.
5. Discard the supernatant, wash the pellet once with 70% ethanol (do not disturb the pellet during the addition of 70% ethanol) and centrifuge for 10 min as in step 4.
6. Discard the supernatant and air-dry the pellet for 2–3 min.
7. Resuspend the pellet in 10.5 μL of water. The pellet should be easily soluble.

### 3.1.5 First-Strand cDNA Synthesis

1. Add 1 μL of random hexamer primers (3 μg/μL) into the sample from Subheading 3.1.4.
2. Incubate at 65°C for 5 min; snap cool on ice.
3. In the meantime prepare the following mix:

| Reagent | Volume (μL) per sample |
|---|---|
| 5× first-strand buffer | 4 |
| 100 mM DTT | 2 |
| 10 mM dNTP mix | 1 |
| RNaseOUT (40 U/μL) | 0.5 |

4. Add the mixture of step 3 (7.5 μL) to the test tube containing the mRNA sample.
5. Mix well and incubate at 25°C for 2 min.
6. Add 1 μL of SuperScript II (200 U/μL), mix by repeated pipetting.
7. Incubate in a thermal cycler as follows: 10 min at 25°C, 50 min at 42°C, 15 min at 70°C, and then 4°C until the next step. If a thermal cycler without a heating bonnet is used, centrifuge the reaction tubes to collect any condensate before proceeding to the next step.

#### *3.1.6 Second-Strand Synthesis*

RNaseH is used to partially digest the template RNA. The RNA fragments are then used as primers to initiate the synthesis of the second DNA strand by DNA polymerase I. Since we deal with DNA from here on, the following reagents need not be DEPC treated.

1. Place the test tubes from Subheading 3.1.5 on ice and add 61 μL of ice-cold water.
2. Add 10 μL of 10× second-strand buffer and 3 μL of 10 mM dNTP mix.
3. Incubate on ice for 5 min.
4. Add 1 μL of RNaseH (2 U/μL) and 5 μL of DNA polymerase I (10 U/μL).
5. Mix gently by repeated pipetting and incubate at 16°C for 2.5 h.
6. Purify the resulting double-stranded DNA either using a Qiagen PCR spin kit or solid-phase reversible immobilization (SPRI) as described below (see Subheading 3.3.2) (see Note 9).

### *3.2 Single-Sample Library Preparation*

In this section we describe the preparation of libraries for RNA sequencing using the cDNA obtained in Subheading 3.1.6 and the adapters designed and marketed by Illumina. This technique of library preparation can be considered when exceedingly high numbers of reads are desired for a given sample. When the library is prepared using the following method, a single sample library is sequenced per lane in an Illumina flow cell. For applications such as differential expression and transcriptional phenotype analysis, a sufficient number of reads can result from pooling of multiplexed

samples, which saves considerable time and money. Preparation of a multiplexed library is described in Subheading 3.3.
We recommend performing all the reactions detailed below with a positive control DNA sample along with the cDNA sample from Subheading 3.1.6. The positive control helps determine the success of the library preparation. It can be 500 ng of a specific 200–300 bp DNA fragment from a PCR reaction dissolved in 10 mM Tris–HCl, pH 8.5. The positive control DNA should be generated with plain (unmodified) primers.

### 3.2.1 End Repair

1. The purified cDNA from Subheading 3.1.6 should be eluted in 30 μL of EB solution.
2. Prepare the following reaction mix:

| Reagent | Volume (μL) per sample | Final concentration in 100 μL reaction |
|---|---|---|
| Water | 27 | |
| 5× T4 DNA ligase buffer | 20 | 1× |
| 10 mM ATP | 10 | 1 mM |
| 10 mM dNTP mix | 4 | 0.4 mM |
| T4 DNA polymerase (3 U/μL) | 3 | 0.09 U/μL |
| Klenow DNA polymerase (5 U/μL) | 1 | 0.05 U/μL |
| T4 polynucleotide kinase (10 U/μL) | 5 | 0.5 U/μL |

3. Add 70 μL of the reaction mix to 30 μL of the purified cDNA and mix by finger flicking the microcentrifuge tube.
4. Incubate at 20°C for 30 min.
5. Purify the end-repaired DNA with a QIAquick PCR spin column and elute with 32 μL of EB solution.

### 3.2.2 Addition of a Single A Base

1. Prepare the following reaction mix:

| Reagent | Volume (μL) per sample | Final concentration in 50 μL reaction |
|---|---|---|
| 10× Klenow buffer | 5 | 1× |
| 1 mM dATP | 10 | 0.5 mM |
| Klenow DNA polymerase (5 U/μL) | 3 | 0.33 U/μL |

2. Add 18 μL of the reaction mix into the 32 μL of end-repaired DNA from Subheading 3.2.1.
3. Incubate at 37°C for 30 min.
4. Purify the resulting DNA with a QIAquick MinElute column, and elute in 24 μL of EB solution.

### 3.2.3 Adapter Ligation

1. Prepare the following reaction mix (see Note 10):

| Reagent | Volume (μL) per sample | Final concentration in 50 μL reaction |
|---|---|---|
| Water | 10 | |
| 5× DNA ligase buffer | 10 | 1× |
| Adapter oligo mix | 1 | |
| DNA ligase (1 U/μL) | 5 | 0.1 U/μL |

2. Add 26 μL of reaction mix to the microcentrifuge tube containing 24 μL of DNA from Subheading 3.2.2 and mix by finger flicking.
3. Incubate at room temperature for 15 min.
4. Purify the adapter-ligated DNA with a QIAquick MinElute column and elute in 15 μL of EB solution.

### 3.2.4 Gel Purification

1. Prepare a 2% agarose gel in 1× TAE buffer such that the thickness of the gel is about 0.5 cm. Include ethidium bromide in the gel.
2. Load 15 μL of the sample from Subheading 3.2.3 next to a well containing 100 bp DNA ladder (see Note 11). For handling multiple samples, leave at least 2 blank wells between samples to prevent cross contamination.
3. Run the gel at 100 V until the 100 bp and 200 bp bands of the DNA ladder are well separated.
4. Cut a gel slice at 200 bp ± 25 bp and purify the cDNA with a QIAquick gel extraction kit.
5. Elute cDNA in 30 μL of EB.
6. Dilute the positive control DNA sample in 75 μL of EB.
7. Prepare a 2% agarose gel in 1× TAE buffer such that the gel thickness is about 0.5 cm. Include ethidium bromide in the gel.
8. Load 30 μL of the diluted positive control DNA next to 150 ng of positive control DNA that has not been subjected to library preparation.
9. Load the 100 bp ladder in a separate well, run the gel as in step 3. Successful reactions should result in a 70 bp increase in the size

of the treated positive control DNA due to adapter ligation. Alternatively, run the positive control samples on an Agilent Bioanalyzer DNA 1000 chip (see Note 12).

#### 3.2.5 PCR Enrichment

1. Set up the following PCR reaction mix and aliquot a 20 μL portion into a PCR tube:

| Reagent | Volume (μL) per sample | Final concentration in 50 μL reaction |
|---|---|---|
| Water | 7 | |
| 5× Phusion HF buffer | 10 | 1× |
| PCR primer 1.1 | 1 | |
| PCR primer 2.1 | 1 | |
| 25 mM dNTP mix | 0.5 | 0.25 mM |
| Phusion DNA polymerase | 0.5 | 0.02 U/μL |

2. Add 30 μL of the DNA from the Subheading 3.2.4 and mix by repeated pipetting.
3. Incubate with the following PCR program: 30 s at 98°C; 15 cycles of 10 s at 98°C, 30 s at 65°C, and 30 s at 72°C; a final extension cycle of 5 min at 72°C.
4. Purify the resulting DNA with a QIAquick PCR spin column and elute in 30 μL of EB solution.
5. Prepare a 2% agarose gel containing ethidium bromide in 1× TAE such that the thickness of the gel is about 0.5 cm and load 25 μL of PCR-enriched positive control DNA next to 30 μL of the remaining positive control DNA obtained after adapter ligation and 150 ng of the original positive control DNA. Include a well containing 100 bp DNA ladder (see Note 11).
6. Run the gel at 100 V until sufficient resolution is obtained between 100 and 200 bp of the 100 bp ladder. A distinct shift in the positive control DNA size should be visible compared to the adapter-ligated positive control DNA after PCR enrichment.
7. Analyze 1 μL of the PCR-enriched DNA on an Agilent Bioanalyzer DNA 1000 chip to assess the quality of the final product and to determine the DNA concentration. Successful preparations should yield a distinct band at ~200 bp. This material is processed further for cluster generation on the Illumina Cluster Station using the manufacturer's recommended protocol.

### 3.3 Multiplexed Library Preparation

In this section we describe a multiplexing technique in which up to 228 samples can be pooled into one lane for sequencing. We adopted and standardized this method for transcriptomic sequencing from ref. (10), which was originally described for pooling

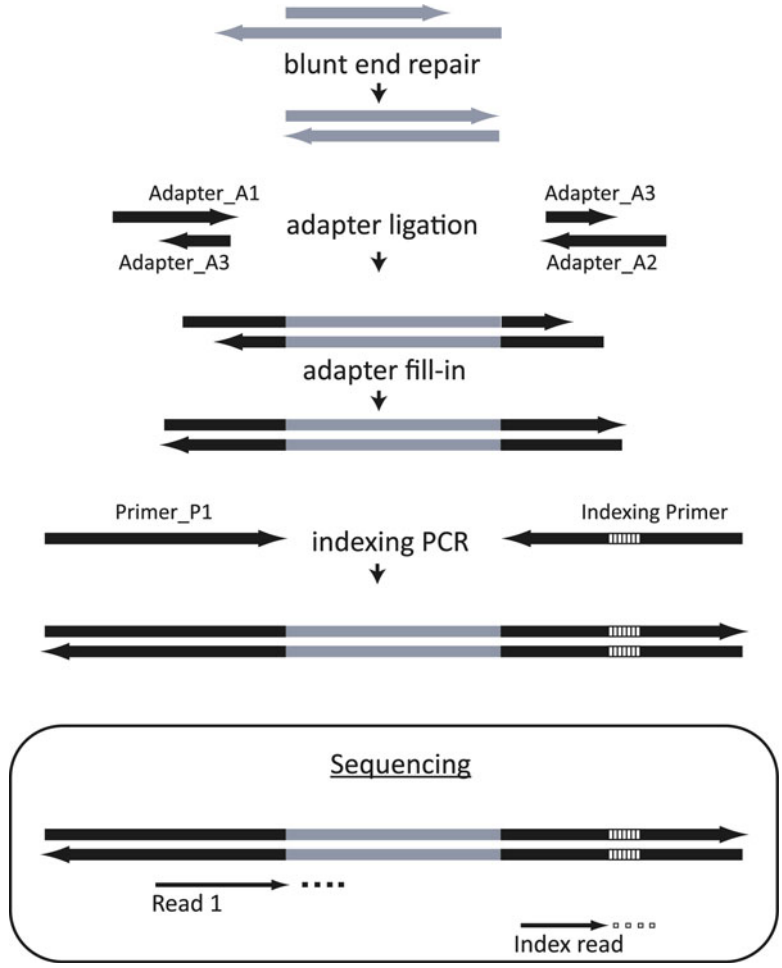


**Fig. 1** A strategy for preparing a multiplexed sequencing library. *Lines* indicate DNA strands. *Gray* indicates the target DNA molecules to be sequenced and *black* indicates adapters. The adapters are ligated to the ends of the target DNA molecules and filled in to make them blunt-ended. Indexing is performed at the last step of library PCR amplification. The indices are depicted as striped segments within the adapters. In the sequencing reaction, they are identified in a separate short sequencing run (Index read) after the initial sequencing of the DNA (Read 1) (Adapted from ref. (10))

genomic samples. The overall strategy of library preparation is outlined in Fig. 1. In this method, DNA barcodes that label unique samples are attached to one of the adapters. Barcoding is performed at the final step of indexing PCR amplification. These barcodes are identified in a separate short sequencing run after the sequencing of the actual cDNA. We have successfully performed as many as 24-fold multiplexing. Pooling fewer than 4 libraries is not recommended (10).

*3.3.1 Preparation of Adapter Mix*

The following reaction produces adapter mixes that are sufficient for 500 samples.

1. Assemble the following hybridization reactions in separate PCR tubes for each hybridization mix:

| Reagent | Volume (μL) | Final concentration in 100 μL reaction |
|---|---|---|
| Hybridization mix for adapter A1 (200 μM): | | |
| Adapter_A1 (500 μM) | 40 | 200 μM |
| Adapter_A3 (500 μM) | 40 | 200 μM |
| Oligo hybridization buffer (10×) | 10 | 1× |
| Water | 10 | |
| Hybridization mix for adapter A2 (200 μM): | | |
| Adapter_A2 (500 μM) | 40 | 200 μM |
| Adapter_ A3 (500 μM) | 40 | 200 μM |
| Oligo hybridization buffer (10×) | 10 | 1× |
| Water | 10 | |

2. Mix the contents by repeated pipetting.
3. Incubate the reactions in a thermal cycler with a heating bonnet for 10 s at 95°C, followed by a ramp down from 95°C to 12°C at a rate of 0.1°C/s.
4. Combine both reactions to obtain a ready-to-use adapter mix (100 μM each adapter). Adapters can be aliquoted into 4 tubes, stored at −20°C, and thawed repeatedly for subsequent use.

### *3.3.2 Reaction Cleanup Using Solid-Phase Reversible Immobilization*

Purify the cDNA from Subheading 3.1.6 using carboxyl-coated magnetic beads (SPRI beads) as explained below. The given ratio of the SPRI beads to the volume of DNA solution is ideal for DNA molecules above 150 bp. The size cutoff for the cleanup reactions can be controlled by varying the amount of beads (refer to the manufacturer's protocol). A 25 bp DNA ladder may be used as a control to standardize the purification protocol. This procedure can be performed using 96-well plates or individual microcentrifuge tubes, depending on the application. A magnetic apparatus suitable for tubes, such as a DynaMag™-2 magnet, should be used in place of a magnetic plate if individual microcentrifuge tubes are used.

1. Resuspend the stock solution of SPRI beads by vortexing. Add 0.05% Tween 20 to the suspension to facilitate subsequent pipetting.
2. Add the SPRI bead suspension to the reactions as follows, using wide orifice pipette tips.
   (a) Add 1.8 volumes of the SPRI bead suspension to each cDNA sample.

(b) Seal the wells with caps and vortex for several seconds. Ensure that the beads are uniformly suspended.

(c) Let the plate stand for 5 min at room temperature.

(d) Collect the liquid to the bottom of the wells by brief centrifugation in a plate centrifuge at 800×*g*. Avoid cross contamination while opening and closing the caps.

3. Place the plate on a 96-well magnetic plate, and let it stand for 5 min to separate the beads from the solution. Discard the supernatant without removing the beads.
4. Leave the plate on the magnetic rack, add 150 μL of 70% ethanol to wash the beads, wait 1 min and then remove the supernatant.
5. Repeat step 4.
6. Remove residual traces of ethanol using a multichannel pipette. Allow the beads to air-dry for 20 min at room temperature without caps.
7. Elute as follows:

(a) Add 30 μL of EBT to the wells and seal the plate with caps.

(b) Remove the plate from the magnetic rack and resuspend the beads by vortexing.

(c) Wait 1 min and then collect the liquid in the bottom of the wells by briefly centrifuging the plate at 800×*g*. The beads may become clumpy but this appearance does not affect DNA recovery.

(d) Place the plate back on the 96-well magnetic plate, wait 1 min, and transfer the supernatant to a new 96-well reaction plate. Carryover of small amounts of beads will not adversely affect subsequent reactions.

### 3.3.3 End Repair

We recommend performing all the reactions with a positive control DNA sample and a negative control along with the cDNA sample from Subheading 3.3.2. The positive control DNA will help determine the success of library preparation. It can be 300 ng of any DNA of about 200–300 bp dissolved in 10 mM Tris–HCl, pH 8.50. If produced by PCR, the positive control DNA should be generated by *Taq*-DNA polymerase with unmodified primers and purified as in Subheading 3.3.2. The negative control is 30 μL of EB solution.

1. Prepare the following reaction master mix for the required number of reactions:

| Reagent | Volume (μL) per sample | Final concentration in 50 μL reaction |
|---|---|---|
| Water | 10.8 | |
| Buffer Tango (10×) | 5 | 1× |
| dNTPs (25 mM each) | 0.2 | 100 μM each |
| ATP (100 mM) | 0.5 | 1 mM |
| T4 polynucleotide kinase (10 U/μL) | 2.5 | 0.5 U/μL |
| T4 DNA polymerase (5 U/μL) | 1.0 | 0.1 U/μL |

2. Add 20 μL of the reaction mix into 30 μL of each cDNA sample from Subheading 3.3.2.
3. Mix the solutions thoroughly by repeated pipetting using a multichannel pipette. Avoid vortexing after the addition of enzymes.
4. Incubate at 25°C for 15 min followed by incubation at 12°C for 5 min.
5. Clean up the reaction using SPRI beads as in Subheading 3.3.2 and elute the end-repaired DNA in 20 μL EBT solution.

### 3.3.4 Adapter Ligation

1. Prepare a master mix of adapter ligation reagents for the required number of reactions. Pipette PEG using a wide orifice pipette tip. Vortex the reaction mix containing all the reaction ingredients before adding the enzyme, to mix the viscous PEG. Dissolve any white precipitate in the ligase buffer by vortexing before adding it to the reaction mix. If the amount of template DNA is higher than 100 ng, increase the amount of adapter mix to 1 μL.

| Reagent | Volume (μL) per sample | Final concentration in 40 μL reaction |
|---|---|---|
| Water | 10.6 | |
| T4 DNA ligase buffer (10×) | 4 | 1× |
| PEG-4000 (50%) | 4 | 5% |
| Adapter mix from Subheading 3.3.1 (100 μM each) | 0.4 | 1 μM each |
| T4 DNA ligase (5 U/μL) | 1 | 0.125 U/μL |

2. Add 20 μL of master mix to 20 μL of end-repaired DNA from Subheading 3.3.3.
3. Incubate at 22°C for 30 min.
4. Clean up the reaction using SPRI beads as in Subheading 3.3.2 and elute with 20 μL of EBT solution.

### 3.3.5 Adapter Fill-In Reaction

1. Prepare a master mix for the required number of reaction as shown below.

| Reagent | Volume (μL) per sample | Final concentration in 40 μL reaction |
|---|---|---|
| Water | 14.1 | |
| ThermoPol reaction buffer (10×) | 4 | 1× |
| dNTPs (25 mM each) | 0.4 | 250 μM each |
| *Bst* polymerase, large fragment (8 U/μL) | 1.5 | 0.3 U/μL |

2. Add 20 μL master mix to the adapter-ligated DNA from Subheading 3.3.4.
3. Incubate at 37°C for 20 min.
4. Clean up the reaction using SPRI beads as in Subheading 3.3.2 and elute with 20 μL of EBT solution.

### 3.3.6 Library Quality Control and Characterization

1. Prepare a 2% agarose gel in 1× TAE buffer such that the thickness of the gel is about 0.5 cm. Include ethidium bromide in the gel.
2. Load 10 μL of the treated positive control DNA next to the original positive control DNA to verify the success of the library preparation reactions. Also load the 10 μL of negative control DNA. Include a well containing 100 bp DNA ladder (see Note 11).
3. Run the gel at 100 V until sufficient resolution is obtained between 100 and 200 bp of the 100 bp ladder. Successful library preparation will cause the positive control DNA size to shift by 67 bp (see Note 13). We recommend carrying over the positive control DNA through the next step of indexing PCR and running another 2% gel after the final step. Expect to see a further 36 bp shift in the DNA size after incorporation of the index oligonucleotides.

### 3.3.7 Library Quantification

Quantify the library by measuring the DNA concentration by quantitative PCR. We recommend using a commercially available quantitative PCR kit containing SYBRE green. This step will ensure equal representation of samples during pooling for multiplexed sequencing.

1. Use a previously quantified indexed library, if available, as a positive control. Dilute this positive control sample in TE buffer to yield an adequate range of concentrations in order to quantify samples that are at least twofold on either side of the probable library concentration. We recommend a range of $10^{-8}$ to $10^{-14}$ g/μL.
2. If no such library is available, positive control DNA from Subheading 3.3.5 can be amplified using indexing PCR primers

as in Subheading 3.3.8 and purified as in Subheading 3.3.2. Determine the DNA concentration of the positive control using a spectrophotometer.

3. Use 1 μL of the library for quantification in a 30 or 50 μL reaction condition. Use 1 μL of positive control DNA at different dilutions as mentioned in step 1 in a 30 or 50-μL reaction for producing a standard curve to quantify the samples. Amplify the library, the positive control, and the negative control using Primer P1 and one of the indexing primers. Use 60°C as the annealing temperature during the quantitative PCR cycle.

The negative control mentioned in Subheading 3.3.3, which is processed along with the positive control DNA through every step of the library preparation, should yield at least twofold less DNA than the library samples. The positive control library DNA can be used to measure the degree of DNA carryover from previous reactions and purifications.

### 3.3.8 Indexing PCR and Sample Pooling

Use equal amounts of DNA from each sample for the indexing PCR. A small portion of the sample DNA is sufficient since the number of amplification cycles can be altered to suit the amount of starting material. We usually perform PCR using 0.1 to a 10 ng of template DNA. This strategy allows saving template DNA in case the indexing PCR reaction fails with the current barcode and a different barcode has to be chosen. Run positive control DNA side by side with the original positive control DNA and pre-indexed positive control DNA to test the success of the library preparation reactions.

1. Prepare the master mix for a sufficient number of reactions:

| Reagent | Volume (μL) per sample | Final concentration in 50 μL reaction |
|---|---|---|
| Water | 37.1– *A* | |
| Phusion HF buffer (5×) | 10 | 1× |
| dNTPs (25 mM each) | 0.4 | 200 μM each |
| Primer_P1 (10 μM) | 1 | 200 nM |
| Phusion Hot Start High-Fidelity DNA Polymerase (2 U/μL) | 0.5 | 0.02 U/μL |
| Add separately to each well | | |
| Indexing primer (10 μM) | 1 | 200 nM |
| Template DNA (library) | *A* | |

2. Add the master mix to each well and perform PCR with the following temperature profile: initial denaturation at 98°C, 30 s; denaturation at 98°C, 10 s; annealing at 60°C, 20 s; elongation at 72°C, 20 s; and final extension at 72°C, 10 min.

The number of cycles that would result in a plateau of the PCR reaction can be determined from the quantitative PCR step in Subheading 3.3.7. Alternatively, adjust the cycle number depending on the template DNA concentration as follows: ≥100 ng: 10 cycles; ≥10 ng: 12 cycles, ≥1 ng: 15 cycles, ≥100 pg: 18 cycles.

3. Clean up the reaction using SPRI beads as in Subheading 3.3.2 and elute the indexed DNA in 40 μL of EB buffer (see Note 14).
4. Remove any leftover magnetic beads before pooling of the samples.
5. Quantify the indexed library. Performing quantitative PCR is the best way to quantify the indexed library, but spectrophotometric quantification may suffice. Pool equal quantities (100–300 ng) of library DNA from each sample. Analyze 1 μL of the pooled product on an Agilent Bioanalyzer DNA 1000 chip to assess the quality of the final product and to quantify the DNA concentration. All the samples should yield similar DNA concentrations at the end unless there was a significant difference in fragment size between the samples.

This material is processed further for cluster generation on the Illumina Cluster Station using the manufacturer's recommended protocol. Most laboratories (including ours) submit their materials to a core facility for Illumina sequencing. This material is ready for submission to the sequencing service for the Illumina sequencing procedure.

### 3.4 Multiplexing: Simulation and Empirical Results

Transcriptome profiling data can be used for investigating multiple patterns of individual gene expression as well as a molecular phenotyping tool (5, 12). The vast amounts of data produced by each sequencing run may sometimes exceed the need, especially for molecular phenotyping and for the analysis of transcript abundance. Multiplexing allows processing of many samples in one sequencing run, thus reducing the cost per sample. The assumption in multiplexing is that the loss of information is uniform across all genes, but we were not sure whether the *Dictyostelium* transcriptome, with its uniquely high A to T content, may behave differently. We tested this assumption by simulations and empirically.

We first analyzed the potential effect of multiplexing by simulation on previously published non-multiplexed data. We then performed a direct experiment with 24-fold multiplexing, which matched our experimental needs, using the RNA samples that were used to obtain the non-multiplexed data. The non-multiplexed dataset was obtained by collecting RNA samples at 4-h intervals during the 24-h developmental program in two independent replicates in *D. discoideum*, and the mRNA samples were analyzed using RNAseq (5). To calculate the similarity between the transcriptional profiles at different time points, we performed hierarchical clustering on the

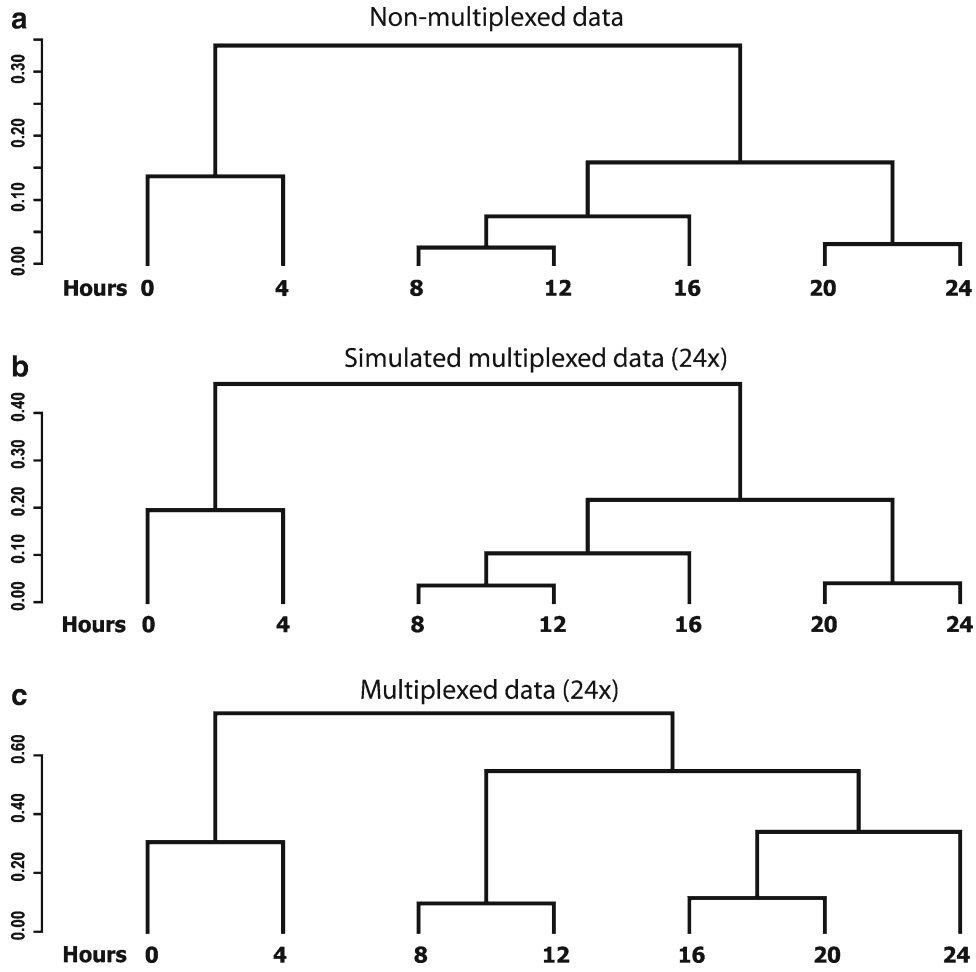


**Fig. 2** Simulated and empirical multiplexing results. The dendrograms depict the distances between the transcriptional profiles at each of the time points (hours). (**a**) Samples analyzed by RNAseq without multiplexing (5). (**b**) Simulated data at 24× multiplexing. Simulation was performed on the data used to generate *panel a*. (**c**) Samples analyzed by RNAseq with 24× multiplexing. The RNA samples used to generate the data for panel a were multiplexed 24-fold and sequenced

expression vectors consisting of all the genes from each time point and visualized the results as a dendrogram (Fig. 2) (5). The expression vectors from each of the time points were scaled to one million counts of all the polyA$^+$ genes, averaged between the two replicates and log transformed to minimize the effects of outliers. We used Pearson's correlation (PC) to calculate the distance ($D = 1 - \mathrm{PC}$) and complete linkage as the clustering criterion. Two objects (individual time points or joints) are joined by a horizontal line if they are more similar to one another than to any other object in the dataset. The vertical distance between objects is inversely proportional to the

similarity between them. The horizontal distances in the dendrogram are meaningless.

We simulated multiplexed data by assuming equal loss of information from all samples. We performed hierarchical clustering on the simulated multiplexed data and observed that the structures of the dendrograms obtained were essentially identical to those obtained with the non-multiplexed data up to 512-fold multiplexing. Figure 2b shows the similarity between time points of the simulated multiplexed data with 24-fold multiplexing. Though there is no theoretical limit to multiplexing 512-fold, the protocol allows only up to 228-fold multiplexing.

For the empirical test, the mRNA samples that were previously analyzed without multiplexing were analyzed with 24-fold multiplexing. We performed hierarchical clustering on the multiplexed data and visualized the similarities between the different time points using dendrograms. We observed that the structure of the empirical data (Fig. 2c) was similar to that obtained when no multiplexing was done (Fig. 2a). The only exception was clustering of the 16-h sample with the 8–12-h clade in the original and simulated data, whereas the 16-h sample was clustered with the 20–24-h clade in the empirically multiplexed data. In either case, the temporal order of the time points was correct. These results indicate that multiplexing does not introduce systematic errors into the data.

As the sequencing technology is improving regularly, we are currently able to obtain more data from each one of the multiplexed samples than we were able to obtain from a single sample in the non-multiplexed method just 2 years ago (5). In the future we may be able to increase the fold of multiplexing further.

## 3.5 Software Tools

It is nearly impossible to provide a complete protocol for analyzing RNAseq data because the methods vary with the research needs. We therefore provide a few examples of routine analyses and the tools we use to perform them.

### 3.5.1 Input Data

The pipeline's principal input is next-generation sequencing (NGS) reads in QSEQ or FASTQ format:

```
Line1    @1
Line2    GAGACCCTCTACAATTCAATGAAAAAGATTTTAGCTTTACCAGAGGATGT
Line3    +
Line4    bbbeeeeeggggiihiagcgiighhdggffhiiaefgcc¢ebghffhii
```

where line1 is sequence identifier, line2 is raw nucleotide sequence, line3 is sequence identifier/description, and line4 is quality values. If reads are different from the reference data used by the pipeline,

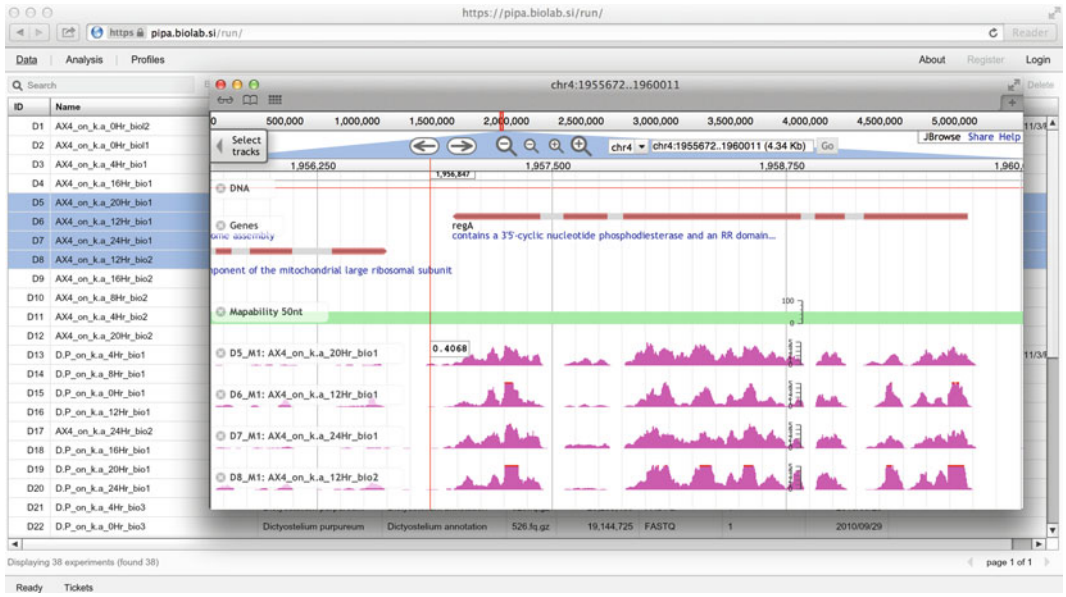


**Fig. 3** PIPA. A view of a list of experiments with raw data, mapping information, and gene expression (*background grids*) and a display of the RNAseq read distribution for a selected dataset (*center*)

they can be complemented with the sequence of the reference genome in FASTA format:

```
Line1   >Chromosome_1
Line2   TTTGGTACAAATGGTTTAACTTCTTCTGGCATACGAAGAGCAATTTCACC…
Line3   >Chromosome_2
Line4   GTTCAAGAAGCCAAACAACAAACCGGCGCTAATGCCACAGTTATTTATGT…
```

and genome annotation (gene features with their locations in GTF format, e.g., the position of 3 exons, gene DDB_G0267698):

| Source | Type | Start | Stop | Strand | Gene id | Transcript id |
|---|---|---|---|---|---|---|
| dictyBase | exon | 624027 | 624219 | - | DDB_G0267698 | DDB0305284 |
| dictyBase | exon | 623830 | 623910 | - | DDB_G0267698 | DDB0305284 |
| dictyBase | exon | 623530 | 623627 | - | DDB_G0267698 | DDB0305284 |

### *3.5.2 PIPA: A Dictyostelium RNAseq Data Management Pipeline*

PIPA (http://pipa.biolab.si, Fig. 3) is a web-based software tool for NGS data management and bioinformatics analysis. Its main task is to manage, map, and preprocess the data. PIPA supports data storage and management, experiment annotation and bioinformatics analysis including de-multiplexing, sequence mapping, estimation of transcript abundance, differential expression analysis, and quality control. It uses a server-based architecture, in which the data analysis

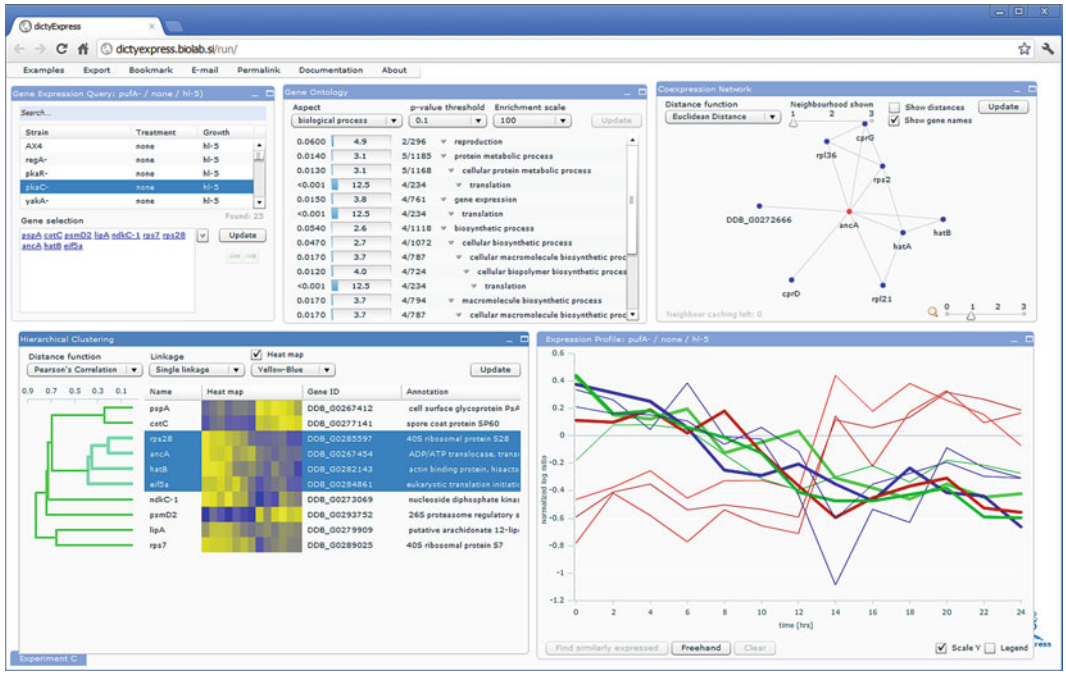

**Fig. 4** dictyExpress. Experiment selection (*top left*), enrichment analysis (*top center*), co-expression network display (*top right*), hierarchical clustering (*bottom left*), and display of gene expression profiles (*bottom right*)

runs on the server and the results are rendered in an interactive web-client with a graphical user interface. PIPA employs standard bioinformatics procedures and implementations, such as FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), Bowtie (13), and Bioconductor (14). The results (mapped reads, counts, and transcript abundance) can be either downloaded and analyzed by a third-party program or analyzed in dictyExpress (15) or Orange (16), which can access the data directly.

### 3.5.3 dictyExpress: Web-Based Gene Expression Analytics

The web-based interactive gene expression analysis program dictyExpress (http://dictyexpress.biolab.si) can query PIPA and render either public or proprietary gene expression data. Its analytics toolbox (Fig. 4) includes visualization of expression profiles, enrichment analysis of Gene Ontology (GO) terms, hierarchical clustering, search of co-expressed profiles, and navigation through gene co-expression networks.

### 3.5.4 Orange with a Bioinformatics Add-On: A Visual Programming Suite for Gene Expression Data Analysis

Orange (http://orange.biolab.si, Fig. 5) is a general-purpose interactive data analytics environment, where data flow schemas can be built from computational units called widgets. Gene expression analysis is implemented through the bioinformatics add-on. The bioinformatics widgets implement various data analysis and visualization tasks, including gene selection, enrichment analysis, exploration of KEGG pathways (http://www.genome.jp/kegg), and access to publicly available data such as Biomart (www.biomart.org) and GO (17).

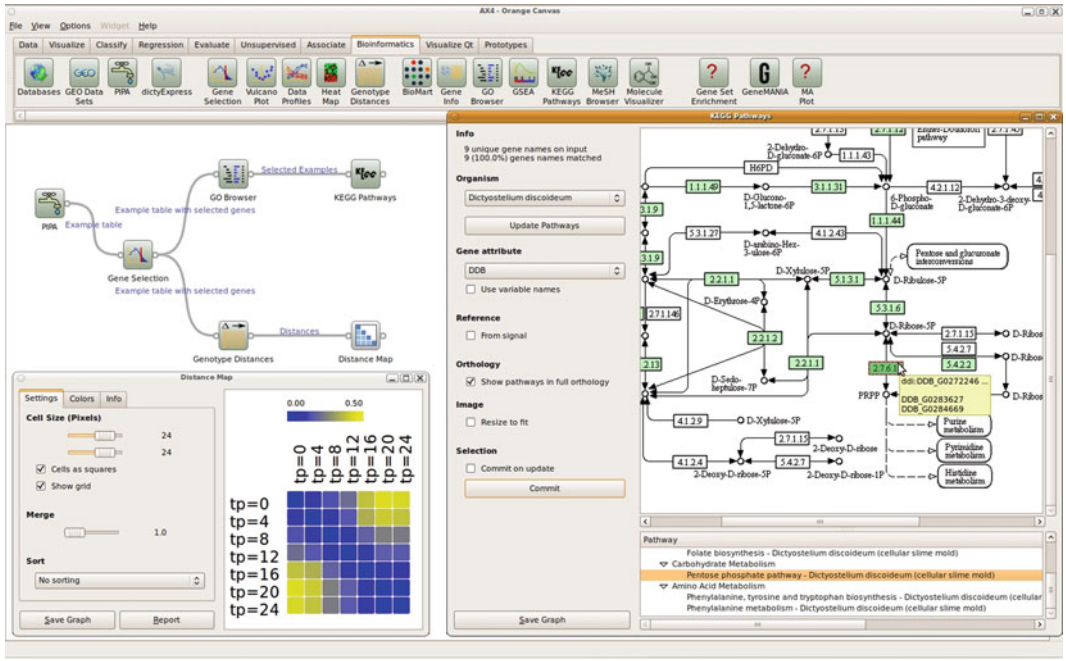


**Fig. 5** A typical Orange bioinformatics schema. Wild-type *Dictyostelium* gene expression data from PIPA are fed to the "Gene Selection" widget. The selected genes are analyzed for term enrichment in Gene Ontology, where a subset of genes is chosen and for which a KEGG pathway is displayed. The other branch of the schema computes and displays differences between expression profiles at different stages of development

Each Orange widget accepts input data and provides output results. Widgets can also interconnect with other visualization, network exploration, and data-mining widgets from the Orange data-mining toolbox to compose sophisticated data analysis schemas.

## 3.6 Data Processing

### 3.6.1 Data Input and Management in PIPA

1. Login to PIPA, go to the "Run PIPA" link.
2. Upload raw sequence data in FASTQ/FASTA format from a local data file or specifying a remote server address (using the "Upload" button in the Data pane).
3. De-multiplex the data, if required (Library pane, "De-multiplex" button).

### 3.6.2 Annotation in PIPA

1. Select a single experiment in the Data pane and click "Edit."
2. Choose an annotation format (e.g., Dictyostelium) and populate the field values (e.g., Experiment name, Time point, Species).
3. If a new field is required, edit the annotation format in "Settings/Annotation formats." Add the desired field (select its type: string, number, date) and position it in the field list.

### 3.6.3 Data Mapping in PIPA

1. Select experiments and initiate mapping to the chosen reference genome. Select the desired mapping parameters and features (e.g., iterative trimming of reads from the 3′ end).

2. Explore mapping statistics including the number of uniquely mapped reads (single-hits) ($N_{UNIQUE}$), the number of unmapped reads ($N_{NOTMAPPED}$), and the number of reads with multiple mappings ($N_{MULTIPLE}$). The system computes three alternative expression values for each gene:
   (a) $Exp_{RAW}$—the number of reads uniquely mapped to gene exons
   (b) $Exp_{RPKM}$—raw gene expression scaled by exon length (reads per kilobase of exon model per million mapped reads) where:

   $$Exp_{RPKM} = 10^9 \times Exp_{RAW} / (N_{UNIQUE} \times Exon_{LENGTH})$$

   $Exon_{LENGTH}$ = length of gene exons (nt)

   $N_{UNIQUE}$ = total number of all uniquely mapped reads from the experiment, excluding the non-polyadenylated genes
   (c) $Exp_{MAP}$—same as $Exp_{RPKM}$, but scaled by the uniquely mappable part of the exons: all possible subsequences of the reference genome (of the same length as reads in raw data) are mapped back to the reference genome, and $Exon_{MAPPABLE}$ is the number of uniquely mapped sequences to the exons

   $$Exp_{MAP} = 10^9 \times Exp_{RAW} / (N_{UNIQUE} \times Exon_{MAPPABLE})$$

3. Explore gene expression values for individual experiments, view read alignments together with gene features (exons, coverage) in jbrowse (http://jbrowse.org/) or download BAM files (includes all mapping results).

## 3.7 Data Analysis

### 3.7.1 Differential Expression Analysis in PIPA

1. Create a new differential expression study by clicking on "Analysis/New/Differential expression."
2. Select experiments for condition A and condition B and choose the analysis method.
3. Differentially expressed genes are shown in the results grid.

### 3.7.2 Expression Analyses in dictyExpress

dictyExpress contains 7 interconnected components. Selecting a gene in one of the components highlights it in all the others and pressing the "Update" button in any component results in propagation and commitment of the selected set in all the other components. We describe three options for exploration, but there are many other ways to select and analyze genes or groups of genes.

#### Searching for Genes by Name

1. In the "Gene Expression Query" component, select an experiment in the upper window (e.g., *D. discoideum* strain AX4 grown on *K.a.*).
2. In the same component, enter the desired gene names in the "Gene selection" window. An interactive menu allows gene selection from a list.

3. Press the "update" button to propagate the selection to the other components.
4. Use the green arrow buttons to return to previous selections.

Searching for Genes by Expression Pattern

1. In the "Expression Profile" component, press the "Freehand" button.
2. Place the brush cursor in the graph window and draw the desired pattern of expression.
3. Press the "Freehand" button again and then press the "Find similarly expressed" button. A new window will appear with gene names.
4. Select the desired genes (up to 30) and press update.

Selecting Differentially Expressed Genes

1. In the "Prespore/Prestalk Differential Expression Analysis" component, select the desired comparison (the default is *D. discoideum* prespore cells vs. prestalk cells).
2. Each spot in the volcano plot represents a gene. The *x*-axis shows the $\log_2$ of the ratio between the selected samples and the *y*-axis represents the degree of confidence. Select a few spots of interest by pointing and clicking on a spot or on a group of spots.
3. Select up to 30 genes from the pop-up box and press the "Update" button.

### *3.7.3 Accessing PIPA Data in Orange*

1. Select the "Bioinformatics" tab in Orange.
2. Place the PIPA widget (Fig. 6) on the canvas and open it by double clicking. The default settings access our published data. To access private data, provide your PIPA user name and password at the bottom left corner.
3. Select the expression type (optional) and specify the type of data transformation. Choose "Average Replicates" to output gene-wise median among replicates, and "Logarithmic Transformation" to log-transform gene expressions (gene expression $x$ is transformed to $\log_2(x+1)$).
4. Select experiments. You may use the "Search" window to find experiments that match terms such as name, species, and strain. After selection, click the "Commit" button to initiate data transfer from the server and place the data in the widget output (see Note 15).
5. Optionally connect the output of the PIPA widget to the input of a "Data Table" widget. The Data Table shows gene expression values with the experiment labels on top and gene IDs in the rightmost column (Fig. 6) (see Note 16).

### *3.7.4 Quality Control in Orange*

1. Load the expression data without replicate averaging (see above). Connect the expression data to the "Quality Control" widget.

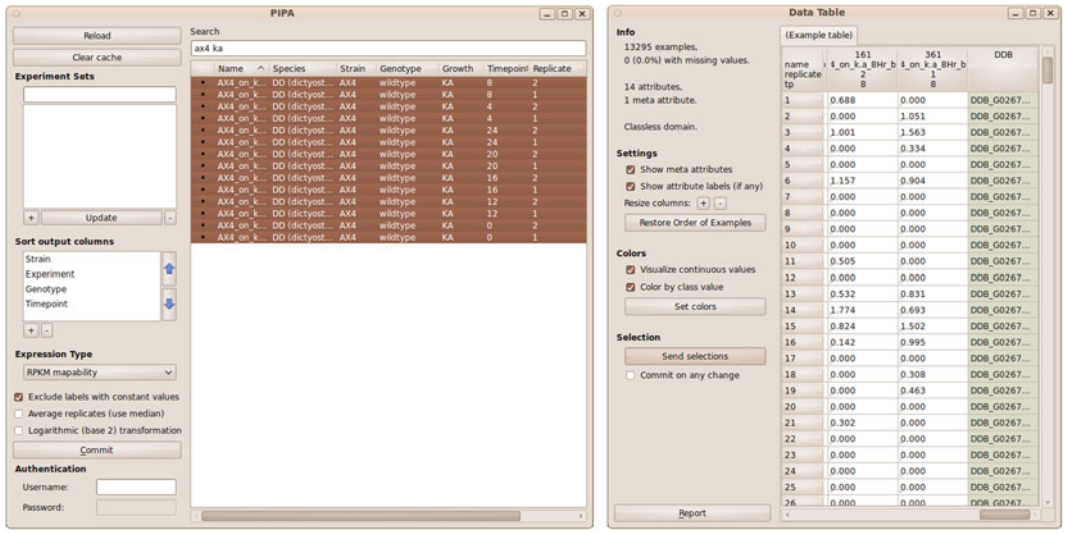


**Fig. 6** Gene expression data selection with PIPA. Data selection and downloading with the PIPA widget (*left*). The gene expression data for the selected experiments are shown in the Data Table widget (*right*)

2. In "Quality Control," select the labels shared by experiments in an experimental group (e.g., wild type or a specific mutant) and a distance metric to compute the gene expression profile distances between different replications of the experiment groups.
3. Explore the results. The widget shows distances between one instance (a reference) in the experiment group and the other instances of the same group. For comparison, distances to all other experiments are visualized as well (Fig. 7). Double clicking on one of the experiments changes it to be the reference.
4. Intuitively, replicates of the same experiment should appear closer to each other than to replicates outside the group. Clear outliers indicate irreproducible samples that can be removed from further analysis either by deselecting them in the PIPA widget or by choosing reproducible experiments in the "Select Attributes" widget. Such experiments should also be annotated accordingly in PIPA (see Subheading 3.6.1).

### *3.7.5 Gene Expression Data Analysis in Orange*

Estimation of Gene Expression Profile Distances or Distances Between Genes

1. Connect the expression data (e.g., from the "PIPA" widget) to the "Attribute Distance" widget to compute distances between expression profiles or to the "Example Distance" widget for distances between genes.
2. Select a distance measure in the distance widget (e.g., "Euclidean distance," "Spearman correlation," "Pearson correlation").

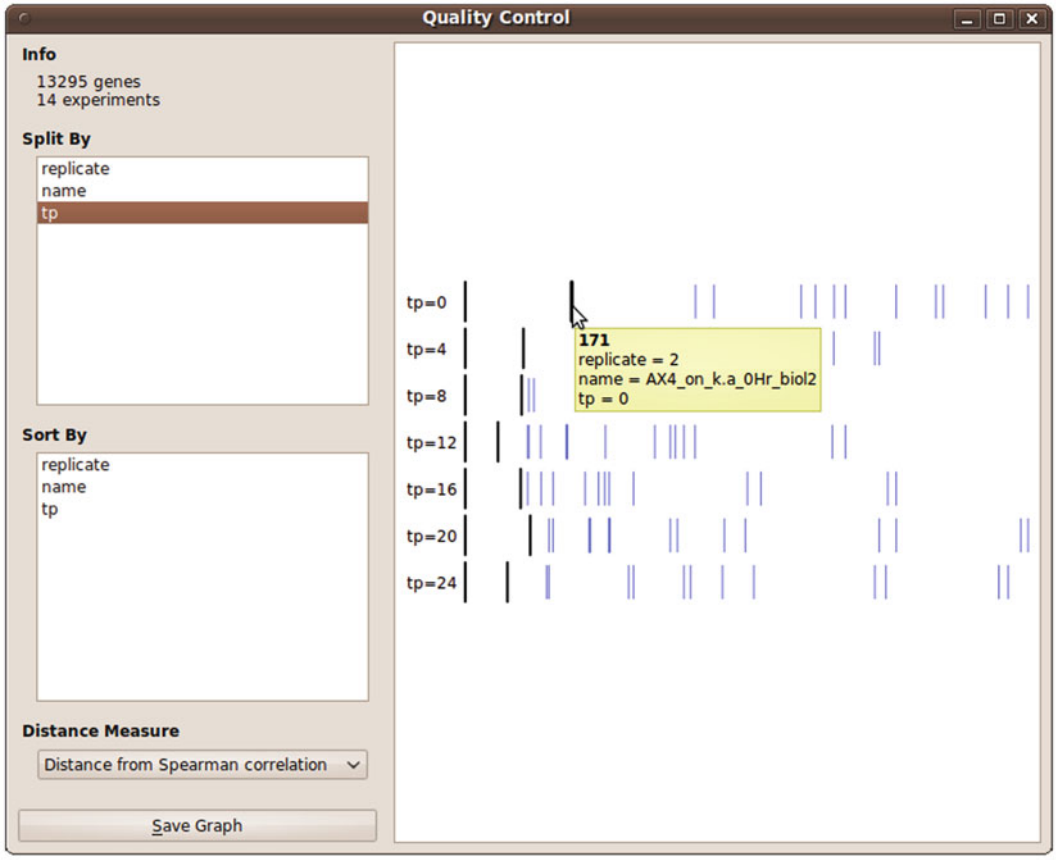


**Fig. 7** Quality control widget in Orange. The tooltip shows the experiment labels

3. Connect the output of the distance widget to one of the widgets for visualization of distances (e.g., "Distance Map," "Hierarchical Clustering," "MDS").

Estimation of Genotype-Specific Gene Expression Profile Distances

1. Connect the expression data (e.g., from the "PIPA" widget) to the "Genotype Distances" widget.
2. In the "Genotype Distances" widget, select the labels shared by the experiments in a group with the same genotype and labels by which to sort experiments within groups. Select a distance metric. Press "Compute" to initiate distance estimation.
3. Visualize the distances as in Subheading 3.7.5.1, step 3.

Gene Ontology Enrichment Analysis

1. Connect the expression data (e.g., from the "PIPA" widget) to the "Gene Selection" widget, which enables gene selection based on differential expression. Differentially expressed genes can also be selected with the "Volcano Plot" widget. In the "Volcano Plot" widget, select target labels and then select genes on the graph. The marked genes will appear in the widget output.

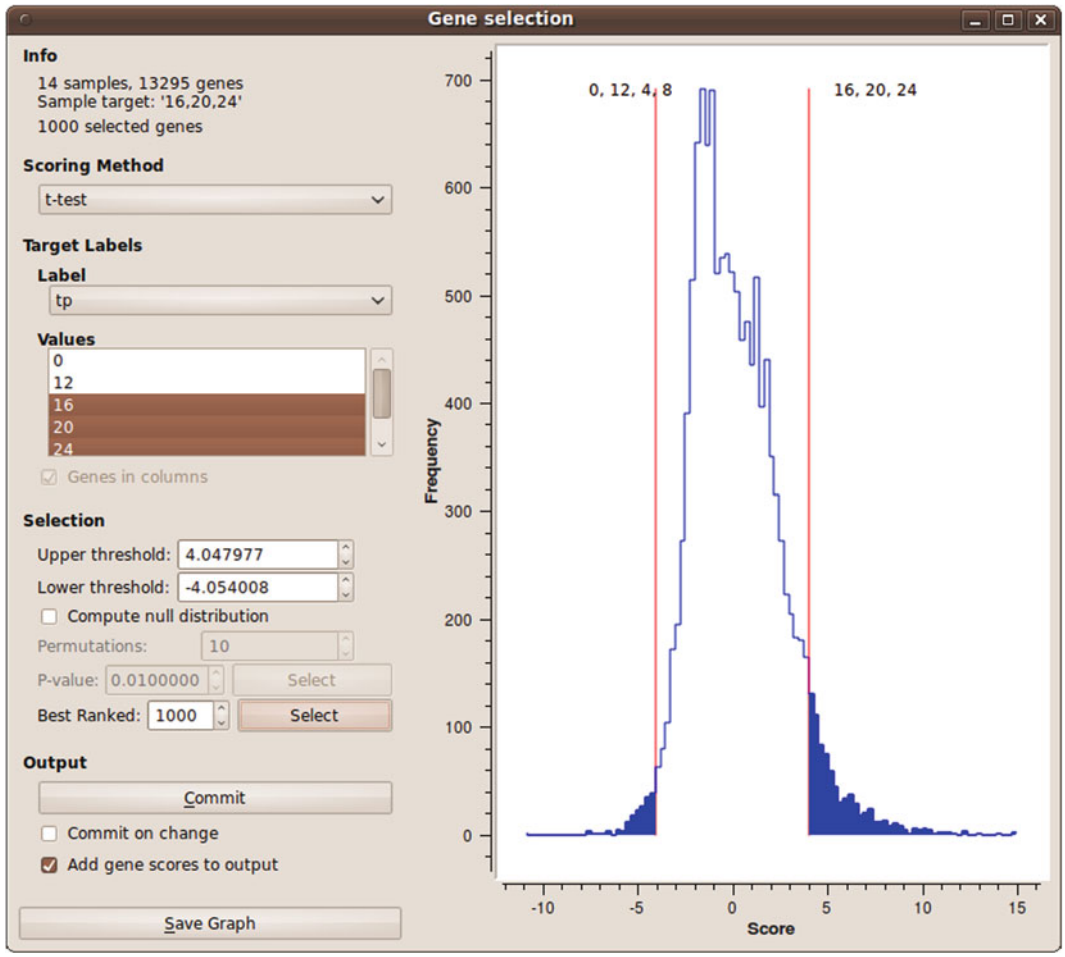


**Fig. 8** The Gene Selection widget in Orange. The highest-ranked 1,000 genes are selected

2. In the "Gene Selection" widget (Fig. 8), select the scoring method and the target labels. After the scores are computed, a histogram is shown in the widget main area.
3. Choose a cutoff point, either according to the *p*-value obtained from permutation tests or specify a number of highest-ranked genes. Click the "Commit" button to send the data to the widget output.
4. Connect the "Gene Selection" to the GO Browser. In the GO Browser, choose the correct organism and select a GO aspect to analyze. If no GO terms are found, increase the *p*-value or reduce the term size threshold ("Filter" tab). The default reference set for the computation of enrichment includes all the genes of the given organism. To use a custom reference set, connect a customized reference set of Genes to the "Reference" input of the GO Browser and choose the "Reference set (input)" option.

5. The enriched terms are displayed in a tree. You may select a term for further analysis. Expression data with genes from the selected terms will appear in the widget output. Analyze these genes through either "Gene Info," "KEGG," "Data Table," or other Orange widgets.

Gene Set Enrichment Analysis

1. Connect the expression data to the "GSEA" widget. Unlike the GO Browser, this widget does not require a preselected subset of genes.
2. In the "GSEA" widget, choose experiments that belong to the two groups you want to compare. Select gene sets in the "Gene sets" tab and click "Compute."
3. A list of enriched gene sets is displayed. Choose a gene set for additional analysis.

Visualization of Distances with a Distance Map

1. Connect the output of "Attribute Distance," "Example Distance," or "Genotype Distances" to the "Distance Map" widget (e.g., Fig. 5, lower left).
2. Observe the distances. Optionally sort the items and display the results of clustering. If an area in the distance map is selected, the widget outputs the respective data subset.

Visualization of Distances with Multidimensional Scaling

1. Connect the output of "Attribute Distance," "Example Distance," or "Genotype Distances" to the "MDS" widget for multidimensional scaling (18).
2. In the "MDS" widget, run optimization (click "Optimize"). Adjust the view in the "Graph" tab (see Note 17).

Hierarchical Clustering of Experiments

1. Connect the expression data to the "Attribute Distance" or "Genotype Distances" widget and select the appropriate settings.
2. Connect the output of the "Distance" widget to the "Hierarchical Clustering" widget (Fig. 9). In "Hierarchical Clustering," set the "Linkage" to "Ward's" and the "Annotation" to "label."

## 4 Notes

1. It is advisable to dedicate an area of the laboratory and a set of pipettors to RNA work. The work area and the pipettors should be cleaned with RNAseZap (Ambion) or a similar product before each procedure.
2. Plasticware should be sterilized by autoclave in glass beakers covered with aluminum foil and dedicated to RNA work.
3. In principle, other RNA purification procedures may be used as well, but we have not tested their suitability for RNAseq.

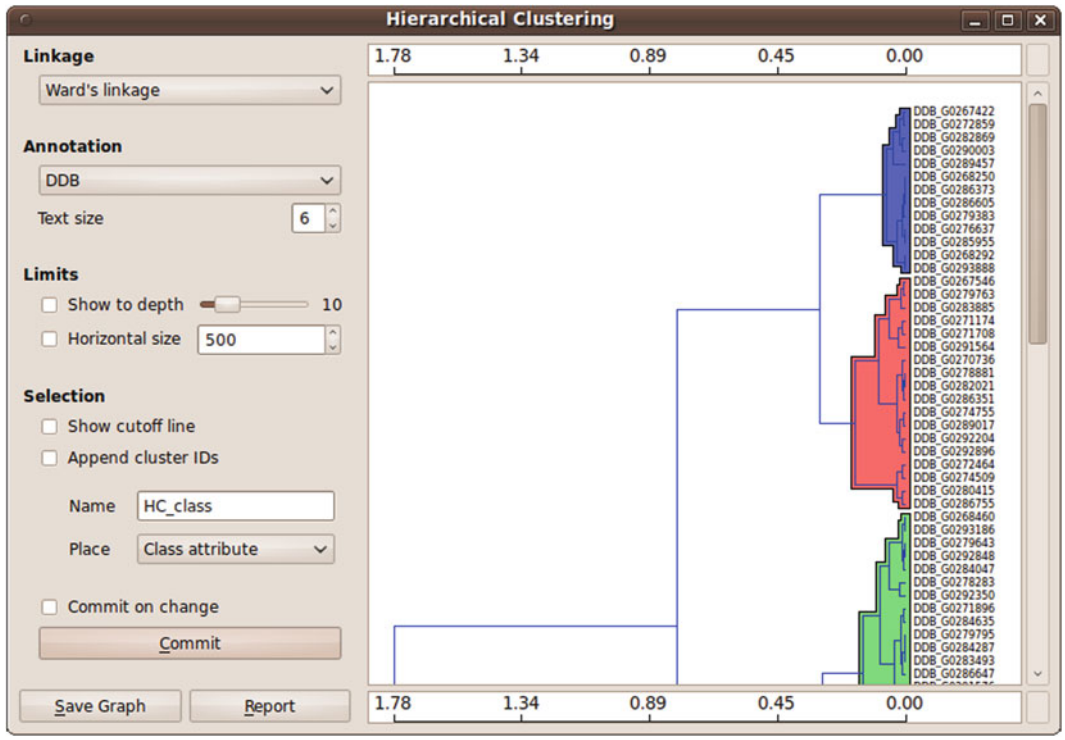


**Fig. 9** Hierarchical clustering of genes and their expression profiles in Orange

4. The RNA is stable in TRIzol for many months under these conditions, and it can be shipped on dry ice if necessary.
5. We have not used smaller or larger amounts of total RNA so we cannot comment on samples outside this range.
6. The procedure can be stopped at this point and the samples can be stored at −80°C for several days. We have not experienced problems with samples stored for as long as 7 days. Thaw the samples to room temperature before the next use.
7. We have successfully used as little as 30 ng of mRNA for library preparation. If one wants to use even lower amounts mRNA from a precious sample, we suggest testing first by comparing the RNA-sequencing results of a comparably small amount of a less precious and more readily available sample to larger amounts of the same mRNA. This analysis would reveal potential skewing in the observed mRNA species abundance.
8. If you wish to fragment the RNA to a different size or to examine the efficiency of fragmentation, we recommend analyzing the samples using an Agilent Bioanalyzer DNA 1000 chip.
9. One may stop at this point and store the samples at −20°C for 2–3 days.
10. The transcriptomic library requires one-tenth the amount of adapters required for a standard genomic DNA library.

11. The DNA ladder should be loaded at 0.1 μg per mm width of well. We usually load 6 μL of the diluted ladder into each well irrespective of the well width.
12. Faint bands of the positive control DNA may indicate loss of DNA through the reactions. In such cases, one may carry the positive control DNA through a PCR enrichment step and run the samples to determine the super shift post-PCR enrichment step. If you do not observe an increase in size after adapter ligation, replace all the enzymes and reagents and try again.
13. If there is no shift in the DNA size, one of the enzymes may have gone bad. Replace all the enzymes and repeat the procedure. If there is no band at all, make sure the SPRI beads are working well. Perform SPRI bead purification of a 25 bp DNA ladder to see the efficiency of the purification. Artifact bands in the negative control indicate cross contamination.
14. The elution is done in EB instead of EBT. In our hands, this gives better readings on the Nanodrop spectrophotometer. Elution with EB may result in some carryover of magnetic beads. This problem can be avoided by collecting only 38 μL of the EB rather than the entire 40 μL.
15. Sets of experiments that are used frequently can be saved.
16. If the Data Table is empty, check if the input is connected to the PIPA widget, and whether there were experiments selected and the "Commit" button clicked in the PIPA widget.
17. If the optimization algorithm is stuck in a local minimum, click the "Jitter" button, which moves the elements slightly, and click "Optimize" again. The "Randomize" button facilitates a complete restart of the MDS optimization.

## References

1. Kibler K, Nguyen TL, Svetz J, van Driessche N, Ibarra M, Thompson C, Shaw C, Shaulsky G (2003) A novel developmental mechanism in *Dictyostelium* revealed in a screen for communication mutants. Dev Biol 259:193–208
2. Morio T, Urushihara H, Saito T, Ugawa Y, Mizuno H, Yoshida M, Yoshino R, Mitra BN, Pi M, Sato T, Takemoto K, Yasukawa H, Williams J, Maeda M, Takeuchi I, Ochiai H, Tanaka Y (1998) The *Dictyostelium* developmental cDNA project: generation and analysis of expressed sequence tags from the first-finger stage of development. DNA Res 5:335–340
3. Van Driessche N, Shaw C, Katoh M, Morio T, Sucgang R, Ibarra M, Kuwayama H, Saito T, Urushihara H, Maeda M, Takeuchi I, Ochiai H, Eaton W, Tollett J, Halter J, Kuspa A, Tanaka Y, Shaulsky G (2002) A transcriptional profile of multicellular development in *Dictyostelium discoideum*. Development 129:1543–1552
4. Booth EO, Van Driessche N, Zhuchenko O, Kuspa A, Shaulsky G (2005) Microarray phenotyping in *Dictyostelium* reveals a regulon of chemotaxis genes. Bioinformatics 21:4371–4377
5. Parikh A, Miranda ER, Katoh-Kurasawa M, Fuller D, Rot G, Zagar L, Curk T, Sucgang R, Chen R, Zupan B, Loomis WF, Kuspa A, Shaulsky G (2010) Conserved developmental transcriptomes in evolutionarily divergent species. Genome Biol 11:R35
6. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63
7. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe

AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5:613–619

8. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349
9. Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, Nislow C (2010) Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res 38:e142
10. Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010(6):pdb prot5448.
11. Fey P, Kowal AS, Gaudet P, Pilcher KE, Chisholm RL (2007) Protocols for growth and development of *Dictyostelium discoideum*. Nat Protoc 2:1307–1316
12. Van Driessche N, Demsar J, Booth EO, Hill P, Juvan P, Zupan B, Kuspa A, Shaulsky G (2005) Epistasis analysis with global transcriptional phenotypes. Nat Genet 37:471–477
13. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
14. Gentleman R, Carey V, Huber W, Irisarry R, Dudoit S (eds) (2005) Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York
15. Rot G, Parikh A, Curk T, Kuspa A, Shaulsky G, Zupan B (2009) dictyExpress: a *Dictyostelium discoideum* gene expression database with an explorative data analysis web-based interface. BMC Bioinformatics 10:265.
16. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B (2005) Microarray data mining with visual programming. Bioinformatics 21:396–398
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
18. Torgerson WS (1952) Multidimensional scaling: I theory and method. Psychometrika 17:401–419